ARI Research Note 92-65

Criterion-Referenced Testing: A User's Resource

J. Douglas Dressel and Angelo Mirabella
U.S. Army Research Institute

Automated Instructional Systems Technical Area
Robert J. Seidel, Chief

Training Systems Research Division
Jack H. Hiller, Director

July 1992

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

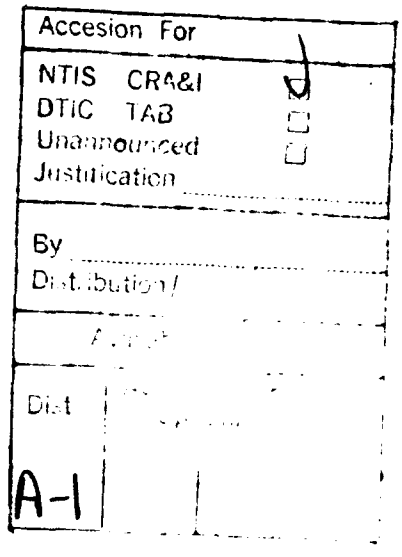
A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

MICHAEL D. SHALER
COL, AR
Commanding

Technical review by

Judith E. Brooks
Joseph D. Hagman



NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1992, July	Final Oct 90 - Oct 91

4. TITLE AND SUBTITLE

Criterion-Referenced Testing: A User's Resource

5. FUNDING NUMBERS

63007A
794
3304
H2

6. AUTHOR(S)

Dressel, J. Douglas; and Mirabella, Angelo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-II
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8. PERFORMING ORGANIZATION REPORT NUMBER

ARI Research Note 92-65

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

--

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

--

11. SUPPLEMENTARY NOTES

--

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release;
distribution is unlimited.

12b. DISTRIBUTION CODE

--

13. ABSTRACT *(Maximum 200 words)*

To provide Army trainers and test-development personnel with basic information and guidance on the rationale and development of criterion-referenced tests (CRTs), a literature review was performed on issues relating to CRT development over the past 15 years. The results of this search were divided into eight topical areas, and an overview of essential principles and an annotated bibliography were written for each of these topical areas. Test developers can use the indicated references for a detailed explanation of issues to address and procedures to follow to produce a valid CRT.

14. SUBJECT TERMS

Criterion-referenced testing
Test development
Military testing
Performance assessment

15. NUMBER OF PAGES

26

16. PRICE CODE

--

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), in cooperation with the U.S. Army Training and Doctrine Command (TRADOC) and with TRADOC's schools, conducts research to develop ways to achieve cost-effective training for the Army. In 1987 ARI joined with the Quartermaster School (QMS) at Fort Lee to identify and solve enlisted supply department training problems. The partnership was defined by a letter of agreement entitled "Establishment of a Joint Training Technology Transfer Activity (TTTA)."

This report is one result of that partnership. The work was carried out by members of the Automated Instructional Systems Technical Area (formerly the Logistics Training Technologies Technical Area) of ARI's Training Systems Research Division to supply QMS and other TRADOC school personnel with a resource for producing criterion-referenced tests of job performance.

CRITERION-REFERENCED TESTING: A USER'S RESOURCE

EXECUTIVE SUMMARY

Requirement:

To provide Army trainers and test-development personnel with basic information and guidance on the proper use and development of criterion-referenced tests (CRTs). Also, to provide these personnel with a means to acquire detailed information on CRT development and applications in military settings.

Procedure:

A literature review of recent findings in CRT issues, construction, and application during the past 15 years was performed. These findings were divided into eight topical areas, and a brief overview of essential principles was written for each area. A bibliography cross-referenced to the topical areas was constructed. Brief annotations were written for journal articles or conference papers that had either military relevance or reported a military application of CRT.

Findings:

A concise resource that generates an awareness of vital CRT application issues and developmental procedures was produced. It directs the user to over 40 recent, topic-specific, source documents.

Utilization of Findings:

Test developers can use this document to construct a CRT. They can also consult the indicated references for a detailed explanation of issues to address and procedures to follow to produce a valid CRT.

CRITERION-REFERENCED TESTING: A USER'S RESOURCE

CONTENTS

CRITERION-REFERENCED TESTING: A USER'S RESOURCE

INTRODUCTION

The purpose of this paper is to provide information on criterion-referenced tests (CRTs) and their development. The paper is primarily intended for trainers or test developers who have limited experience with CRT development. The paper has an extensive annotated bibliography that can be of benefit to more experienced testing personnel, especially those with an interest in CRT testing in military situations. This bibliography will also benefit those readers who wish a fuller and more technical explanation of the principles and concepts of CRT development.

Structure of This Document

The paper consists of two sections: the topical overviews and the bibliography. The topical overviews present a summary of the principal findings, concepts, and issues that a test developer should understand when considering the construction of a CRT. The CRT topics reviewed are

a. General Definitions and Descriptions: provides a basis for distinguishing between norm-referenced tests and CRTs.

b. Test Construction: provides information on the various developmental steps involved in test construction.

c. Content Domain Specification: provides a framework by which to determine the scope and content of the test. Various approaches to this determination are considered.

d. Test Item Writing: provides guidelines for the creation of individual test items.

e. Item Analysis: notes the diagnostic value of item analysis for both the developed test and the preceding instruction.

f. Standard Setting: explores the issues and the rationales for establishing different criterion values to match the function of the test.

g. Reliability: defines the roles of reliability in criterion-referenced testing and how reliability can be measured.

h. Validity: provides a framework for determining the validity of a CRT.

Following the topical overview section is the bibliography. The bibliography represents the culmination of computer-searches of various scientific databases for the period of 1975 through 1990. Also included are conference papers selected from the proceedings

of the annual conferences of the Military Testing Association for the period 1985 through 1990.

How to Use This Document

It is suggested that the user read the overview for a general understanding of the CRT issue, then consult one or more references for a more detailed, technical account of the topic. It is further recommended that the user read a book chapter on the topic prior to reading military applications within the subject area. By using this strategy, the user will be aware of the general issues before examining a specific application that may be narrow in scope.

TOPICAL OVERVIEWS

General Definitions and Descriptions

A student's performance on a CRT indicates how well that student can perform the well-defined objectives of that test. The objectives of the test could cover a task, a duty position, or a period of instruction. The test content can be broad or narrow; the breadth is not important. What is important is the defining of the test objectives which drive the construction of the test. The test objectives form the domain or body of knowledge/skills of the test, which is the criterion to which the test items are referenced. By comparing the student's test score to a minimum standard, the student is either classified as being a "master" (having mastered the content of the test) or a "nonmaster" (lacking the minimal competence required).

The purpose of a criterion-referenced test is to determine if the student has mastered the content area (domain) of the test. For example, assume the test is referenced to a criterion of job skills which form a military occupational specialty (MOS). If the student passes this test, then the assumption is made that he/she has mastered the MOS. No assumption can be made as to how this student compares to other students within the class. This is a concern of norm-referenced tests which have a different function from criterion-referenced tests. Basically a norm-referenced test will indicate how well a student performed the test as compared to the performance of his/her classmates. A criterion-referenced test will indicate if the student can perform the job (objectives) for which the test was developed.

Test Construction

Construction of a criterion referenced test is a relatively straight forward operation. While each step must be considered and performed with care, the developer must also consider the relative importance of the test against the time and effort to be spent on its development. For example, constructing a test measuring the students' performance on a simple hour of instructional material would not require the same rigorous development as an end-of-course test. The commonly accepted test construction steps are presented below. Descriptions of these steps are provided in later sections of this overview.

1. Define the purpose of the test
2. Review the individual objectives
3. Draft test items to fit the objectives
4. Ensure a review of test items by content (SMEs) and test specialists
5. Edit test items
6. Perform a tryout (field test)
7. Revise test items
8. Assemble test
9. Select standard

10. Pilot test revised test
11. Prepare administration manuals
12. Collect task item statistics

Content Domain Specification

Content domain specification refers to determining which content areas, subjects, or tasks should form the body of the test. Test developers call the field of knowledges/performances covered by the test, the domain of the test. The domain is divided into clearly defined objectives, and individual test items are written for each objective. Content domain specification basically concerns detailing exactly what the test should test.

Testing for competence for a specific job requires the performance of a front-end analysis. This process would require that observers record the actual tasks the job-holders perform and note the features of acceptable performance during the completion of each task. This group of tasks would therefore define the job. A panel of highly knowledgeable people in this job area could select tasks which are critical for successful job performance. These critical tasks could then be sampled by constructing test items in a uniform/standardized manner for the test.

Developers of tests for an Army military occupation specialty (MOS) have some of this work done for them. Training and Doctrine Command (TRADOC) has lists of critical tasks recorded for each MOS. The test developer could have subject matter experts (SME) examine the list to ensure its current accuracy.

More information will be provided under the heading of Test Item Writing on how to construct individual test items. What is important to remember here is that the domain of the test must be clearly defined and stated. Only in this way can the results of the test be meaningfully interpreted. It must be clear exactly what the successful test-taker has "mastered".

Test Item Writing

There are two principal forms of test items. Student performance on written test items can indicate the student's knowledge underlying task performance while hands-on performance measures can indicate whether the student can perform the task.

Due to physical, time and scoring constraints, most classroom testing employs written tests. There are four common types of written test items. These items are

1) multiple-choice
2) true/false
3) matching
4) constructed response.

By far the most popular written test item is the multiple-choice item. Multiple-choice items consist of a statement or question stem and typically four answer choices or options. There is one correct choice; the remaining three incorrect options are called foils.

Approach the writing of multiple-choice items by following these three general rules:

1) Be sure the question is clearly stated and requires the student to respond.
2) Write the correct choice.
3) Write the foils in the same style as the correct choice.

Some frequent errors in constructing multiple-choice items include:

1) long question stems,
2) grammatically incorrect question stems and options,
3) correct option longer than foils,
4) foils belonging to a set or category different from the correct response,
5) clue in question stem,
6) use of negative/confusing statement,
7) non-random order of correct options.

The test developer can also consider the use of true/false test items. Here a statement is presented and the student judges it to be either true or false. While the probability of correctly guessing a single test item is high (50%) the probability of guessing correctly an entire series of questions is quite low. For example, correctly guessing the answers to twenty true/false items in a test of thirty items would occur about twice in a million occasions.

Follow these general rules when writing true/false items:

1) a single test item should test a single idea or bit of information,
2) make positive statements,
3) avoid long statements,
4) deal with clear-cut facts not disputable issues.

Matching questions are actually a form of multiple-choice question with more than four possible answers. Matching questions can cover a large topical area very efficiently. The rules for writing multiple choice questions also apply to writing matching questions.

Constructed-response items are a different form of question which require the student to recall or create the answer to the question rather than select it from the options presented. There are three forms of constructed-response items:

1) completion: where the student fills in the blanks of a statement,
2) short essay: where the student writes several sentences on the topic questioned,

3) extended essay: where the student writes extensively creating a position drawing upon an entire unit of instruction to answer the question.

Due to scoring constraints, military test developers, especially for enlisted MOSs at the entry-skill level, rarely use the constructed-response item formats.

Hands-on performance testing can either measure the process (the performance) or the product (what was created) to indicate students' competence or skill. Three methods are frequently used to assess skill:

1) observation
2) checklists
3) rating scales

Observation is used when: the student's response is either correct or incorrect, the student either achieved or failed to achieve the objective, e.g., the student bench-pressed 150 pounds or did not. Observation is therefore used when only a single outcome is recorded.

Checklists are used to record the performance on a series of observational responses usually required in a specified sequence.

Rating scales are used to record the performance along a continuum, e.g., from good to bad, high to low. Rating scales are used to rate somewhat abstract qualities or characteristics which may vary gradually. Of these three methods of measuring skill, the creation, use and interpretation of the rating scale requires the greatest care.

In conclusion, it should be noted that regardless of the form, the test item should be developed within context of the domain specification. In order to be of any value, the test item must examine some aspect of the specified topical content.

Item Analysis

The first step in conducting an item analysis is to again review the test items to make certain they reflect the content area you wish to examine. This review should be performed by some content matter expert other than the test developer. Any test items that don't seen appropriate after this review should be deleted.

Next, the draft test items should be field tested with one or more groups of students comparable to those students who will take the actual test. The results of this testing will provide information on how well the items function. Three general testing approaches are commonly used.

1) Preinstruction-postinstruction, in which the same group of students take the test, then receive the instruction, and finally are retested after instruction. Students' item performance is compared before and after instruction.

2) Uninstructed - instructed groups, where two different groups of students receive the test. One group has not received the instruction, while the other group has. Again, the item performance of the groups is compared.

3) Contrasting groups, where the members of each group are individually selected on the basis of either being a master or nonmaster of the content material. The two groups take the test and item performance is compared.

Immediately after field testing the items, the tested students can be asked to provide feedback on the test items. Generally the students are asked:

1) Were there any confusing items?
2) Were there any words in the items which you did not know?
3) Was there any difficulty in understanding what you were asked to do?
4) Were there any items without a correct answer?
5) Were there any items with more than one correct answer?

The student feedback can then be considered and used in conjunction with the item statistics (which follow) to revise the test items.

Item difficulty is the percentage of students who correctly answered the item. The item difficulty index values can range from 0 (an extremely difficult item) to 100 (a very easy item). The difficulty index must be determined for both the instructed and noninstructed students. Index values in the range 0 to 50 would indicate difficult items for the noninstructed students while index values of 70 to 100 would indicate easy items for the instructed students. Difficulty index values can give an indication of the influence of instruction or even the need for instruction.

An item discrimination index refers to how well a test item indicates to which group (instructed, master vs. uninstructed, nonmaster) a student belongs. There are several different forms of this index. However, basically each index operates on the different proportions of students from each group getting the item correct. Index values range from +1.00 to -1.00; a value of +1.00 would indicate that all the instructed students correctly answered the question while none of the uninstructed students' correctly answered the question. A value of +.25 would indicate that 25% more instructed students correctly answered the question than uninstructed students. While test items with high positive discrimination values are preferred, however their selection should not come at the risk of lowering content validity. Only if two test items address the same content area, can the test developer

discard the item with the lower discrimination value without jeopardizing the content validity of the test. Finally, any test item with a negative (-) discrimination value (more nonmasters correctly answering the item than masters) should be examined closely.

Another part of item statistics is the choice response analysis for multiple-choice test items. Here the test developer compares the response pattern for each item from both of the field-tested groups (uninstructed, nonmaster vs. instructed, master). If any of the three following conditions are not met, then the test item probably needs revision.

1) No distractor/foil should receive as many responses from the instructed group as the correct answer.

2) All distractors should receive some (5-10%) responses from each group.

3) Each distractor/foil should be selected by more students from the nonmaster group than from the master group.

Apart from the arena of item statistics, all items should be reviewed to detect any racial, ethnic, sex, or cultural bias. Bias is present if membership in any group would hinder performance regardless of ability.

Standard Setting

A standard is used to classify students as either having mastered a set of objectives or not having mastered those objectives. A standard therefore represents a point on a scale of performance. Scoring above this point indicates competence while scoring below this point indicates a deficiency. While this concept of a magical point of mastery may seem untenable, it is essential for the decision-making role of criterion-referenced tests. There are many methods for standard setting and all require human judgement. The goal of that judgement is to minimize the incorrect classification of students.

Three factors should be considered when setting standards. Briefly these factors are:

1) Analysis of decision context which considers such things as: the consequences of the decision to fail a student, the opportunities for retesting, the availability of remedial training and the consequences of false classification.

2) Clarity of target competencies which allows standard-setters to determine meaningful minimum competence standards. This is a re-iteration of the purpose of the test: it should provide an unambiguous description of the skills being measured.

3) Presence of relevant performance data: refers to pretesting selected groups of students before setting a standard. This would involve administering the criterion-referenced test to groups of uninstructed students, instructed students and previously instructed students. Then from the distribution of group test scores, making a judgement upon where the standard should be set.

Four common approaches to setting standards will be briefly described.

Informed judgment is a method whereby judges set test standards based upon the presence of relevant field test data. A panel of judges arrives at a decision specifying the standard.

Borderline group is a method whereby students, who are thought to be "at or near the borderline" regarding competence of the targeted skill, are selected to take the test. The median test performance of this group then becomes the standard.

Contrasting groups is a method whereby two groups of students are selected to take the test. One group consists of students who are judged and selected (by their trainers) to be clearly masters of the targeted skill, while the other group is composed of students who are selected on the basis of clearly not possessing the targeted skill. Frequency distribution curves for the test performance of two groups are plotted and the point of intersection is used when determining the standard.

A method very similar to Contrasting groups is Criterion groups; here, however intact classes of instructed students act as masters while uninstructed groups act as nonmaster for standard setting purposes.

A final approach to standard setting is Nedelsky's method where the individual test items are evaluated. The multiple-choice test items are judged on the probability of a minimally competent student guessing the correct answer. The correct - by - guessing probabilty of each test item is determined; these item probabilities are then added to generate the total test probability. This could be used as the test standard.

In conclusion, it should be noted that standards should be periodically reviewed. Course content can change, remedial instruction can change, student flow can vary, all of which could prompt a review of the current test standard.

Reliability

Reliability refers to the consistency of measurement. If you carefully measured a board in the morning, upon remeasurement in the afternoon, evening or next year you would have the same measurement. The concept of test reliability is the same. Without any additional instruction, you would expect a student's test score

to be largely the same upon retesting. However, unlike the board measurement, it is highly unlikely that the two test scores will be identical.

Although there are many types of reliability, one is central to criterion-referenced testing. This reliability refers to consistent classification of a student as being either a master or nonmaster of the tested objectives without any additional training/instruction. The simplest form of this index was introduced by Hambleton and Novick (1973). It requires two administrations of the criterion-referenced test to the students without any test feedback. The index is simply the sum of the two proportions of students who received the same master/nonmaster classification after each test. For example, fifty students took the test once and about a week later took the test again. Assume, twenty students passed the test both times and twenty-five students failed the test both times. Then the reliability coefficient (po) equals 20/50 + 25/50 or .90.

Other methods which require only one test administration exist. However these methods both require more complex statistical computations and assumptions which the test developer may wish to avoid.

Validity

The validity of a criterion-referenced test concerns the accuracy with which the scores from the test can be used to achieve the stated purpose of the test. Validity refers to the appropriateness of the decisions which were made, based upon the test results.

Validity does not refer to the test itself but does refer to the use or interpretation of the test score. Validity is never proven conclusively. However, data (the amount of which is a function of the importance of the test) are collected which can indicate whether the test appears to serve its intended function.

There are three principal types of validity of interest to the test developer. These are item, content, and criterion validity.

Item validity involves comparing each individual test item with the domain specifications or objectives of the test. This comparison is performed by a group of SMEs. Any items not clearly matched with an objective are deleted.

Content validity concerns not only item validity but also how well the test item represents the domain/objectives. As a group, the test items should form a representative sample of the specified domain/objectives. Again, a panel of subject matter experts can review the assembled items.

Criterion-related validity concerns how well the criterion-referenced test score/classification predicts future performances. To be meaningful, these future performances should largely represent actual applications of the criterion behaviors about which the test was created. Criterion-prediction data, using various forms of correlation between test scores and performance, can be gathered in experimental or test development evaluation situations. In either case, the initial nonmasters will require remedial training.

However, this would not be appropriate as an ongoing approach to assure the continued validity of the test. This would require that all students advance to: the next stage of instruction, graduation, or development (whatever the setting of the predicted behavior) without regard to their past test performance. This would be injurious to the student's education and could be physically dangerous in some situations.

A more suitable method to determine criterion-related validity which poses fewer hazards, is decision-validity. Decision-validity gives one indication of the accuracy of mastery classification decision based upon test scores and the current standard. Quite simply, it is the sum of the percentages of correct classifications of masters and nonmasters using either the Constrasting groups or the Criterion groups techniques previously described in the standards setting section of this paper.

BIBLIOGRAPHY OF CRITERION-REFERENCED TESTING AND EVALUATION METHODOLOGY

1. General Definitions and Descriptions

Campbell, C. P., & Allender, B. R. (1988). Procedures for constructing and using criterion-referenced performance tests. C.V.A./A.C.F.P. Journal, 23(3), 2-9.

Campbell, C. P., & Hatcher, T. G. (1989, May/June). Testing that is performance-based and criterion-referenced. Performance and Instruction, 1-9.
This article is an overview which describes the rationale and multiple uses of performance-based (hands-on) criterion-referenced testing. A description of the development and validation of these tests is presented.

Cantor, J. A., & Hobson, E. N. (1986). The development of a model for construction of criterion-referenced system achievement tests for the strategic weapon system training program. Paper presented at the 70th Annual Meeting of the American Educational Research Association. [ED 268 178]
This paper describes a consensus/committee approach to the development of content area, item selection, and cut-score determination as a method of CRT development within the context of a naval weapons training program.

Cantor, J. A., & Walker, L. (1985). Criterion-referenced testing for the U.S. Navy's nuclear submarine fleet. Proceedings of the 27th Annual Conference of the Military Testing Association, 832-837.
Discusses the procedure for the development of a criterion-referenced testing program currently in place in the naval submarine fleet.

Hambleton, R. K. (1990). Criterion-referenced testing methods and practices. in T. B. Gutkin & C. R. Reynolds (Eds.) The Handbook of School Psychology. New York: Wiley.

Hambleton, R. K. (1985). Criterion-referenced assessment of individual differences. In C. Reynolds & V. L. Wilson (Eds.) Methodologies and statistical advances in the study of individual differences. New York: Plenum Press.

Hambleton, R. K. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, Winter, 48(1), 1-47.

Nasca, D. (1988). An educator's field guide to CRT development and use in objective-based programs. [ED 293 878]
Author attempts to clarify CRT terminology and offers practitioners a basic view of CRT use and development. An extensive reference list is presented.

Nitko, A. J. (1985). Defining "Criterion Referenced Test". In R. A. Berk (Ed.) A Guide to Criterion-Referenced Test Construction. Baltimore, MD: The Johns Hopkins University Press.

Nitko, A. J., & Pettie, A. (1989). The sixteen quality indicators: standards for evaluating criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Association. [ED 306 293].
Authors present 16 indices developed to assess the quality of SQTs. Indices support guidance provided in Army doctrine for SQT development. Authors feel these indicators can be used in the development of any CRT.

Rudolph, S. A. (1990). Test design and minimum cutoff scores. Proceedings of the 32nd Annual Conference of the Military Testing Association, 204-209.
Paper discusses a structured approach to test design to assure tests are reliable, valid, and of equal difficulty in retest situations. Approach considers criticality of objectives, number of test items and their difficulty, and the establishment of validity and cut-off scores.

Shaycoft, M. F. (1979). Handbook of criterion-referenced testing. New York, NY: Garland Stmp Press.

Schrock, S., Mansukhani, R., Coscarelli, W., and Palmer, S. (1986). An overview of criterion-referenced test development. Performance and Instruction Journal, August, 3-7.
This article presents the major stages in the design and development of CRTs while noting the differences in construction of these tests and norm-referenced tests.

Wimmer, W. D., & VanLandingham, C. W. (1987). Criterion-referenced testing in the US Army Service Schools. Proceedings of the 29th Annual Conference of the Military Testing Association, 509-512.
Paper presents an overview of the importance of testing (CRT) in Tradoc schools/field and provides a listing of military test design references (23) from 1962-1982.

2. Test Construction and Administration

Buck, L. S. (1987). Procedures for the development of trade-skill tests. Proceedings of the 29th Annual Conference of the Military Testing Association, 380-384.

Millman, J. (1984). Individualizing test construction and administration by computer. In R. A. Berk (Ed.) A Guide to Criterion-Referenced Test Construction. Baltimore, MD: The Johns Hopkins University Press.

3. Content Domain Specification

Albert, W. L. (1990). Development of generalized equations for predicting testing importance of tasks. Proceedings of the 32nd Annual Conference of the Military Testing Association, 310-315.
A method was developed by which "testing importance ratings" could be assigned to tasks without the expense of SME-conducted survey. Part of the Air Force's automated test outline (ATO) work for test development.

Baker, G. H., & Laabs, G. J. (1988). Issues in job sample testing. Proceedings of the 30th Annual Conference of the Military Testing Association, 571-575.
Paper discusses a number of practical and technical issues in the development and administration of a hands-on performance test of work samples (Part of the Joint-Service Job Performance Measurement and Enlistment Standard Project.)

Bart, W. M. (1985). How qualitatively informative are test items?: a dense item analysis. Proceedings of the 27th Annual Conference of the Military Testing Association, 707-712.
A psychometric exposition on the definition and attributes of dense test items. A dense test item indicates why students provide incorrect answers and indicates the sequence of remedial instruction.

Buck, L. S. (1987). Procedures for the development of trade-skill tests. Proceedings of the 29th Annual Conference of the Military Testing Association, 380-384.
Describes the procedures developed and implemented for the production of content-valid written and performance tests designed to assess the skill of navy shipyard workers in 17 skill areas.

Distefano, M. K., Pryer, M. W., and Erffmeyer, R. C. (1983). Application of content validity methods to the development of a job related performance rating criterion. Personnel Psychology, 36, 621-631.
Notes the procedure used to: a) develop a content valid set of job requirements and b) application of that set of requirements to evaluate prospective worker's performance.

Dittmar, M. J., Hand, D.K., & Phalen, W. J. (1990). Estimating the importance of tasks by direct task factor weighing. Proceeding of the 32nd Annual Conference of the Military Testing Association, 316-321.

Gifford, J. A., & Hambleton, R. K. (1981). Construction and use of criterion-referenced tests in program evaluation studies. Academic Psychology Bulletin, 3, 411-436.
Technical considerations associated with item selection sampling, and reliability assessment are weighed when using a CRT to evaluate the effectiveness of a program of instruction rather than the performance of an individual.

Laabs, G. L., & Baker, H. G. (1989). Selection of critical tasks for Navy job performance measures. Military Psychology, 1(1), 3-16.
Describes a method for selecting job tasks which when assembled form the content area for work sample performance test. This hands-on performance test would be used as a "benchmark test" for various predictor tests of job performance. The selection of job tasks involved the use of SMEs, job incumbents, and their supervisors.

Maier, M. H. (1985). On the content and measurement validity of hands-on job performance tests. Proceedings of the 27th Annual Conference of the Military Testing Association, 311-316.
Paper examines the content and measurement validity of prototype hands-on performance tests for three Marine Corps specialties. Research is part of the Job Performance Measurement Project.

Mann, W. G. (1989). External validation of job analysis results. Proceedings of the 31st Annual Conference of the Military Testing Association, 205-208.
Author notes the frequent lack of predictive validity of CRTs which are purported to be content-valid. Performed research showing an r=.78 between measures of content and predictive validity.

Phalen, W. J., Albert, W. G., Hand, D. K., and Dittmar, M. J. Estimating testing importance of tasks by direct task factor weighing. Proceedings of the 32 Annual Conference of the Military Testing Association, 316-321.
Paper presents a possible procedure to select tasks for inclusion into an automated task-data-based outline for the development of Air Force Specialty Knowledge Tests (SKTs). Procedure had SMEs rate the importance of each of the seven factors which would then be used to rate specific tasks for possible inclusion into the testbed.

Popham, W. J. (1984). Specifying the domain of content or behaviors. In R. A. Berk (Ed.) A Guide to Criterion-Referenced Test Construction. Baltimore, MD: The Johns Hopkins University Press.

4. Test Item Writing

Millman, J., & Westman, R. S. (1984). Computer-assisted writing of achievement test items: toward a future technology. Journal of Educational Measurement, Summer, 26(2), 177-190.

Roid, G. H. (1984). Generating the test items. In R. A., Berk (Ed.) A Guide to Criterion-Referenced Test Construction. Baltimore, MD: The Johns Hopkins University Press.

Roid, G. H., & Haladyna, T. M. (1982). A Technology for Test-Item Writing. New York: Academic Press.

Vineberg, R., & Joyner, J. N. (1985). Simulation of hands-on testing for Navy machinist's mates. Proceedings of the 27th Annual Conference of the Military Testing Association, 323-326. Discusses the rationale and developmental concerns for constructing a simulated (written) test in lieu of a hands-on performance test.

5. Item Analysis

Kalisch, S. J., Jr. (1989). Use of item response patterns to predict examine performance. Proceedings of the 31st Annual Conference of the Military Testing Association, 163-166. Author presents the case for using item-response patterns on items to increase the efficiency of testing in both adaptive testing (appropriate branching) and non-adaptive testing (termination of session) situations.

Rushano, T., Williams, J. E., and Stanley, P. P. (1990). Item content validity: its relationship with item discrimination and difficulty. Proceedings of the 32nd Annual Conference of the Military Testing Association, 386-391.
Paper describes the relationship between item content validity ratings to item discrimination and difficulty. Part of Air Force Specialty Knowledge Test (SKT) development.

Seddon, G. M. (1987). A method of item-analysis and item-selection for the construction of criterion-referenced tests. British Journal of Educational Psychology, 57, 371-379.
The author presents a method (which uses the basic item statistics of point biserial correlation, mean, and standard deviation) for selecting individual test items from the domain which minimizes sampling error. The theoretical basis and an empirical application are presented.

6. Standard Setting

Arabian, J. M., McHenry, J. J., and Wise, L. L. (1988). Synthetic validation procedures for identifying selection composites and cut scores. Proceedings of the 30th Annual Conference of the Military Testing Association, 434-439.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, Spring, 56 (1), 137-172.
This review provides a trilevel categorization process for 38 methods of standard setting. Ten criteria (Technical/Practical) for method evaluation are presented. The review is intended to facilitate the selection of the suitable method appropriate for the intended application.

Cantor, J. A., & Hobson, E. N. (1986). The development of a model for construction of criterion-referenced system achievement tests for the strategic weapon system training program. Paper presented at the 70th Annual Meeting of the American Educational Research Association. (ED 268-178)

Kulik, C. C., & Kulik, J. A. (1986). Mastery testing and student learning: a meta-analysis. Journal of Educational Technology Systems, 15 (3), 325-345.
Review indicates that end of course scores are highly influenced by criterion level of performance required to progress through course. Implications are evident for setting high cut scores for CRT quizzes within course.

Pettie, A. L. (1985). Standard setting methods for skill qualification tests. Proceedings of the 27th Annual Conference of the Military Testing Association, 391-394.
Notes earlier work on standard setting approaches for the SQT, (largely superseded by author's 1987 MTA paper).

Pettie, A. L., Brittain, C. V. (1987). Establishing minimum pass scores for Skill Qualification Tests. Proceedings of the 29th Annual Conference of the Military Testing Association, 391-394.
Paper notes several approaches to setting minimum passing scores (MPS) on the SQT which has became largely a paper and pencil multiple-choice test. Method used in FY 87.

Rudolph, S. A. (1990). Test design and minimum cutoff scores. Proceedings of the 32nd Annual Conference of the Military Testing Association, 204-209.
Author presents an approach to test design and two procedures to establish test standards. Early results suggest success in Navy training school application.

Walker, C. L., & Cantor, J. A. (1987). Alternative performance standard methodologies: a comparison of results on a strategic weapon system (SWS) missile technical C-R SAT. Proceedings of the 29th Annual Conference of the Military Testing Association, 498-503.
Paper explains the procedures of using three different approaches to set performance standards for a naval weapons training program.

7. Reliability

Arabian, J. M., McHenry, J. J., & Wise, L. L. (1988). Synthetic validation procedures for identifying selection-composites and cut scores. Proceedings of the 30th Annual Conference of the Military Testing Association, 434-439.
Presents a procedure for identifying selection-composites for military occupational skills candidate placement. Notes efficiency in cost savings of analysis and personnel savings as a function of accurate setting of standards.

Kane, M. T. (1986). The role of reliability in criterion-referenced tests. Journal of Educational Measurement, Fall, 23(3), 221-224.
This article focuses upon the importance of a minimal level of internal reliability in a CRT. Generally, any CRT with a reliability of less than .5 should be seriously questioned.

Raju, N. S. (1982). The reliability of a criterion-referenced composite with the parts of the composite having different cutting scores. Educational and Psychological Measurement, 42, 113-129.
The author proposes a method of determining the internal consistency (reliability) for a CRT having subtests with different cut scores. This is not a measure of the reliability of the mastery classification.

Subkoviak, M. J. (1988). A practioner's guide to computation and interpretation of reliability indices for mastery tests. Journal of Educational Measurement, 25(1), 47-55.
This article provides guidance on the meaning, computation, and use of the coefficients of agreement and kappa to determine CRT reliability. Tables are presented which allow these coefficients to be determined from a single test administration.

8. Validity

Brittain, C. V., & Vaughan, P. R. (1987). A comparison of hands-on and written common task test (CTT) scores. Proceedings of the 29th Annual Conference of the Military Testing Association, 385-128.
Notes the relationship between a hands-on performance test of 17 basic soldiering tasks and an alternative paper and pencil test of these tasks. Notes lack of comparability of several task tests.

Buck, L. S. (1989). Are performance tests necessary? Proceedings of the 31st Annual Conference of the Military Testing Association, 123-128.
Focus of research is evaluation of the contribution of written and performance tests to the assessment of a job-incumbents abilities. Used Navy personnel from 17 shipyard trades.

Campbell, C. H., & Campbell, R. C. (1990). Job performance measures for Non-commissioned officers. Proceedings of the 32nd Annual Conference of the Military Testing Association, 541-596. Developed a test battery to measure performance in three job components (supervisory, common, MOS-specific) via three measurement modes (written, hands-on, and ratings). Measurement instruments developed for nine military occupational skills (MOS); intercorrelation of test mode results presented.

Carretta, T. R. (1988). Cross-validation of an experimental pilot selection and classification test battery. Proceedings of the 30th Annual Conference of the Military Testing Association, 559-564.
Use of the basic attributes tests (BAT) in conjunction with the standard Air Force Officer Qualifying Test (AFOQT) will allow a predetermination of flying specialty (fighter vs non-fighter) prior to, rather than following the 52 week undergraduate training program.

Doyle, E. L., Campbell, R. C. (1990). Hands-on and knowledge tests for the Navy radioman. Proceedings of the 32nd Annual Conference of the Military Testing Association, 529-534. Paper notes the development, administration and results of a benchmark hands-on performance tests which would guide the development of written tests which could be used as substitute measures of hands-on job proficiency.

Heneman, R. L. (1986). The relationships between supervisory rating and results-oriented measures of performance: a meta-analysis. Personnel Psychology, 39, 811-826.
Results of a meta-analysis found a low correlation between supervisors' rating of worker performance and objective measures of workers' performance. Author notes limitations of study and advocates the use of composite (multiple rating items) ratings and a relative, rather than absolute, rating format.

Maier, M. H. (1985). On the content and measurement validity of hands-on job performance tests. Proceedings of the 27th Annual Conference of the Military Testing Association, 311-316.
Paper examines the content and measurement validity of prototype hands-on performance tests for three Marine Corps specialties. Research is part of the Job Performance Measurement Project.

Vineberg, R., & Joyner, J. N. (1985). Simulation of hands-on testing for Navy machinist's mates. Proceedings of the 27th Annual Conference of the Military Testing Association, 323-326.
Discusses the rationale and developmental concerns for constructing a simulated (written) test in lieu of a hands-on performance test.

Williams, J. E., Stanley, P. P., & Perry, C. M. (1990). Implementation of content validity ratings in Air Force promotion test construction. Proceedings of the 32nd Annual Conference of the Military Testing Association, 235-240. Paper reviews the historical issues concerning item content validity ratings and integration of these ratings into current test development procedures. This Air Force research is part of the current Specialty Knowledge test (SKT) program.